Bioingeniería Aplicada en el Diagnóstico de Enfermedades

Ana Lía Carabio[1], Elizabeth Silva Layes[1], Fabián Frola[1], Marcelo A. Falappa[2]

[1] Facultad de Ciencias de la Administración
Universidad Nacional de Entre Ríos
Monseñor Tavella 1424 – Concordia – Entre Ríos
anacarfcad.uner.edu.ar,
elizabeth.silva@gmail.com,
fabianenlared@gmail.com

[2] Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Av. Alem 1253 – Bahía Blanca – Buenos Aires
mfalappa@cs.uns.edu.ar

Resumen. El sector salud administra grandes volúmenes de datos, centrándose, la mayoría de las tomas de decisiones en el área clínica. Por tal motivo, contar con información útil, inmediata y efectiva es sumamente relevante en éste ámbito.

En este sentido, la minería de datos es una herramienta que permite encontrar patrones de comportamiento de utilidad para la toma de decisiones clínicas, como lo son la realización de estudios epidemiológicos, cálculo de expectativas de vida, identificación de terapias médicas satisfactorias para diferentes enfermedades, entre otros.

El objetivo del presente trabajo es desarrollar un componente que, integrado a la Historia Clínica Electrónica (HCE), y mediante técnicas de minería de datos, permita servir de apoyo en la toma de decisiones clínicas tanto en el diagnóstico clínico como en la prevención de enfermedades y epidemias.

Palabras claves. Big Data, Minería de Datos, Historia Clínica Electrónica, Sistemas de Soporte a las Decisiones Clínicas.

1 Introducción

En la actualidad, el procesamiento de grandes volúmenes de datos (*Big Data*) para la toma de decisiones ha dejado de ser privativo de organizaciones comerciales y de negocios, y se ha inmiscuido en otros ámbitos, con actividades e intereses diversos. Consecuentemente, el sector salud es uno de los sectores que se ha visto beneficiado con la utilización de herramientas de análisis de datos.

La vasta acumulación de datos clínicos existentes en los *Electronic Health Records* (EHR's) [1] (diagnósticos, tratamientos indicados, paraclínicas, medicamentos suministrados, procedimientos realizados, etc.) presentes en la mayoría de las instituciones sanitarias, brinda una oportunidad inmejorable para:

- La realización de estudios epidemiológicos (análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, entre otros).
- Cálculo de expectativas de vida.
- Identificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.
- Estudio de factores de riesgo para la salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente.
- Identificación de terapias médicas y tratamientos erróneos para determinadas enfermedades [2].

El sector sanitario, en su totalidad, es uno de los que administra los mayores volúmenes de datos, centrándose la mayoría de las tomas de decisiones en el área clínica. Por esta razón, contar con información útil, inmediata y efectiva es sumamente relevante en éste ámbito.

En este sentido, la minería de datos es una herramienta que permite encontrar patrones de comportamiento de utilidad para la toma de decisiones, relacionándose de manera estrecha con la estadística, usando técnicas de muestreo y visualización de datos, depuración y cálculo de indicadores. Según Hand et al. [3], "la minería de datos es el análisis de grandes conjuntos de datos observacionales para encontrar relaciones insospechadas, y para resumir los datos en nuevas formas, comprensibles y útiles para el titular de los datos".

Entendiendo que uno de los campos de aplicación fértiles de la minería de datos es, precisamente el campo de la bioingeniería, nos proponemos integrar el conocimiento que formula la minería de datos a la *Historia Clínica Electrónica* (HCE) como apoyo en la toma de decisiones clínicas.

El objetivo del presente trabajo es desarrollar un componente integrado a la HCE, que permita, a partir del conocimiento adquirido, apoyar las decisiones médicas tanto en el diagnóstico clínico como en la prevención, pudiendo brindar el asesoramiento de forma pasiva -solicitado por el clínico-, o activa -a través de alertas- de acuerdo al tipo de soporte.

2 Minería de Datos

Según Witten et al. [4], la minería de datos es "la extracción de información implícita, previamente desconocida y potencialmente útil de los datos". Y para lograr esto, se construyen programas de computadoras que, automáticamente, buscan en las bases de datos regularidades y/o patrones.

En la minería de datos se involucran diferentes algoritmos para realizar distintas tareas. Estos algoritmos tratan de ajustar el modelo más cercano a las características de los datos a analizar. Los modelos utilizados pueden ser descriptivos o pre-

dictivos [5, 6]. Los modelos descriptivos se utilizan para la identificación de patrones en los datos, aplicando técnicas tales como *clustering*, sintetización/asociación y visualización, entre otras. Los modelos predictivos, en cambio, se utilizan para pronosticar o predecir una situación particular de acuerdo a los datos recopilados de otras situaciones. Clasificación, regresión y análisis de series temporales, entre otras, son algunas de las técnicas del modelado predictivo, que permiten, por ejemplo, hacer un diagnóstico de una enfermedad particular en un paciente, no solamente por su historia, sino también por los resultados de tratamientos en otros pacientes con síntomas similares [5, 6].

En el ámbito médico la aplicación de la minería de datos interesa a varias áreas:

- En el plano clínico resulta de ayuda para la identificación y diagnóstico de distintas patologías. Igualmente, es relevante para el descubrimiento de posibles interrelaciones entre diversas enfermedades.
- A nivel de medicina preventiva, resulta de interés para la detección de pacientes con factores de riesgo para sufrir una determinada patología.
- A nivel de gestión hospitalaria, se puede utilizar para obtener predicciones temporales que permitan optimizar los recursos disponibles y, de ese modo, priorizar el uso de los diversos tratamientos para una misma patología [7].

2.1 Aplicación en el área sanitaria

Como ya se ha mencionado, las tecnologías de la información en el área sanitaria han permitido el registro electrónico de datos, contemplando -entre otros- datos demográficos del paciente, registros sobre el progreso de tratamientos, resultados de paraclínicas, medicación prescripta, diagnósticos, etc. La aplicación de la minería de datos sobre esta información recopilada, presenta, entre otras, las siguientes ventajas:

- Reduce la subjetividad en la toma de decisiones médicas y brinda nuevo conocimiento médico útil para optimizar el cuidado de la salud.
- Colabora con los médicos en la mejora de sus diagnósticos y en la planificación de los tratamientos.
- Permite desarrollar sistemas de control inteligentes mediante el procesamiento de señales biomédicas.
- Optimiza los procesos clínicos en términos de calidad médica y administrativa [6].

Entre los obstáculos que puede encontrar la aplicación de minería de datos en la medicina, podemos mencionar, entre otros:

- la voluminosidad y heterogeneidad de los datos médicos,
- la complejidad de su representación,

- la posible incompletitud de los mismos,
- el alcance y la potencial inconsistencia de dichos datos,
- la no estandarización o no estructuración del lenguaje médico,
- la integración de los mismos,

que conlleva a las instituciones sanitarias a realizar grandes inversiones en tiempo y dinero para poder procesarlos adecuadamente [6].

En cuanto a lo referido al tratamiento de la voluminosidad y heterogeneidad de los datos, se ha visto mejorado por la aparición de las bases de datos NoSQL. En particular, las del tipo orientadas a columnas (*Column-Oriented Databases*), adecuadas para aplicarlas en minería de datos y aplicaciones analíticas, por su forma de almacenamiento, la compresión eficiente de los datos y por su diseño, que permite cargar y analizar grandes cantidades de datos [8, 9, 10].

3 Implementación

Como ya se ha mencionado, el objetivo del trabajo fue desarrollar un componente -del tipo *web service*- que pueda prestar servicios a una HCE y que permita, a partir del conocimiento adquirido, apoyar en la toma de decisiones médicas tanto en el diagnóstico clínico como en la prevención.

Para esto, hemos desarrollado una aplicación en lenguaje Java [11], integrando el software WEKA [12, 13] a dicho desarrollo. Para efectuar la demostración de predicciones se partió de la determinación de un set de atributos relevantes, que permitiera clasificar los distintos casos de manera de generar aprendizaje significativo aplicable a futuras ocurrencias, cotejando, además, distintos algoritmos de clasificación disponibles [14, 15, 16].

En una primera instancia, se trabajó sobre la predicción de diagnóstico de determinados tipos de cáncer, considerando como set de atributos relevantes los datos relacionados con hábitos de vida, datos demográficos, alimentación, antecedentes personales y familiares, entre otros.

Partiendo de la incorporación de estos insumos -que seguramente podrían provenir de distintas fuentes de interés-, hemos pasado a trabajar sobre el aprendizaje. Para esta implementación en particular, el conjunto de datos utilizados para generar el modelo predictivo reside en un repositorio orientado a columnas.

Luego de varios estudios y considerando la amplitud de la parametrización que tienen los algoritmos de clasificación explorados, se pudo determinar aquellos que fueron más relevantes para el campo de acción que se pretendía analizar.

Como técnica de clasificación se utilizaron algoritmos basados en árboles de decisión -especiales para clasificar y predecir-, cuya construcción del clasificador no requiere conocimiento del dominio, el éxito del modelo depende de los datos y tienen la capacidad de construir modelos interpretables [4], [6], [16]. En una primera instancia, se realizaron varios intentos con el algoritmo PART. Este algoritmo produce una lista de decisión sin restricciones, utilizando el procedimiento de división y conquista. Se construye un árbol de decisión parcial de un nivel [4], y se utiliza con datos nominales.

Posteriormente, se trabajó con el algoritmo C4.5: una versión mejorada del algoritmo de árbol de decisión ID3 cuya principal aplicación es el diagnóstico de enfermedades dados los síntomas. Este algoritmo permite trabajar con valores continuos para los atributos y no genera árboles demasiado frondosos [4]. En WEKA, el algoritmo C4.5 se implementa con el nombre J48 [4], [13], [16] y, luego de diferentes análisis, resultó ser el elegido debido a que presentó los mejores resultados en la clasificación de instancias para nuestro campo de aplicación.

Dada la característica de lo que se pretende de la aplicación y del carácter de la muestra, -donde cada clase está codificada en referencia a CIE-10[1]-, se ha incorporado un archivo de propiedades con la interpretación de la codificación. Este archivo muestra el mensaje adecuado, interpretable por el profesional y, además, es factible cotejar con una tabla de referencia que permite una adecuada interpretación del resultado.

4 Caso de estudio

Como caso particular de estudio primario, se seleccionaron 940 pacientes entre 45 y 65 años de edad en una determinada ciudad de la región, captados en el año 2015, con el objetivo de diagnosticar precozmente aquellas personas con probabilidad de adquirir determinados tipos de cáncer.

El conjunto de datos obtenidos para analizar, de acuerdo al objetivo planteado, consta de los siguientes atributos:

1. Sexo: sexo de la persona. Nominal [M, F].
2. Edad: edad de la persona. Entero.
3. Ant_HTA: antecedentes de hipertensión arterial. Nominal [NO, SI].
4. Ant_Diabetes: antecedentes de diabetes. Nominal [NO, SI].
5. Ant_Cancer_Colon: antecedentes de cáncer de colon. Nominal [NO, SI].
6. Ant_Cancer_Mama: antecedentes de cáncer de mama. Nominal [NO, SI].
7. Fumador: si la persona es fumadora. Nominal [NO, SI].
8. Ingesta_Alchohol: si la persona ingiere alcohol. Nominal [NO, SI].
9. Dislipemia: alteración de los niveles de lípidos (grasas) en sangre. Nominal [NO, SI].
10. Diabetes: presencia de diabetes. Nominal [NO, SI]
11. Sobrepreso: tiene sobrepeso. Nominal [NO, SI]
12. Sedentarismo: no realiza actividad física. Nominal [NO, SI]
13. Enfermedades_Mentales: presencia de enfermedad mental. Nominal [NO, SI]
14. HTA: padece hipertensión arterial. Nominal [NO, SI].
15. Nivel_Socio_Económico: nivel socioeconómico del paciente a juicio del profesional, en 4 grupos. Nominal [Alto, Medio, Bajo, Deficitario].

[1] *Clasificación Internacional de Enfermedades*, décima versión.

16. Class: clase de enfermedad que padece o no. Nominal [códigos CIE10].

Considerando puntualmente el ejemplo presentado, si bien el muestreo no fue lo suficientemente discreto careció, además, de atributos de peso que permitieran una predicción positiva de la enfermedad, siendo destacable la necesidad de tener consideraciones respecto a esto. Sin perjuicio de lo detectado, nos comprometimos a continuar con el desafío, teniendo presente que nos movemos en un ámbito muy ajustado.

Observando la disposición presentada por la matriz de confusión generada [Tabla 1], se aprecia que sobre una muestra de 940 casos, sólo un poco más de 30 eran clasificables en clases que determinan una predisposición a casos positivos de enfermedad, distribuidos conforme a CIE-10 para los casos de localización de tumores. Esta clasificación representa un error del 3,4%, resultando clasificadas muy pocas clases o rangos de clases, lo que reduce las posibilidades del aprendizaje.

Tabla 1. Matriz de confusión generada con el algoritmo J48

Matriz de confusión del algoritmo J48								
a	b	c	D	e	f	g	h	<-- classified as
0	0	0	0	0	0	0	1	| a = C07
0	4	0	0	0	0	0	6	| b = C15-C26
0	0	4	0	0	0	0	9	| c = C43-C44
0	0	0	4	0	0	0	3	| d = C50-C58
0	0	0	0	4	0	0	5	| e = C61
0	0	0	0	0	0	0	2	| f = C64
0	0	0	0	0	0	0	1	| g = D04
0	0	0	0	0	0	0	897	| h = NO

Por lo expuesto, fue necesario ajustar la parametrización del algoritmo, sacrificando la confianza de la poda del árbol de decisión, a fin de detener el crecimiento del árbol. Según Mitchell [17] "dado un espacio de hipótesis H, una hipótesis h ∈ H se dice que hay sobreentrenamiento de los datos si existe alguna hipótesis alternativa h' ∈ H, tal que h tiene un error menor que h' sobre los ejemplos de entrenamiento, pero h' tiene un error menor que h sobre la distribución entera de ejemplos".

Así entonces, alejándonos de los valores por defecto que proporciona la librería, resignando la poda, obteniendo árboles de decisión más grandes que consideraban menos nodos mínimos en una rama disminuida a un caso, variando los parámetros del número de pliegues del árbol y considerando a la muestra como conjunto de entrenamiento, se logró una clasificación de un 97,12 % de instancias correctas [Tabla 2], disminuyendo notablemente el error para un árbol de decisión que, en definitiva, no había crecido tanto.

Tabla 2. Clasificación utilizando J48

Clasificación utilizando el algoritmo J48		
Indicador		***Valor***
Correctly Classified Instances	913	97.13%
Incorrectly Classified Instances	27	2.87%
Kappa statistic		0.5353
Mean absolute error		0.0136
Root mean squared error		0.0823
Relative absolute error		56.75%
Root relative squared error		78.04%
Total Number of Instances		940

Cotejando con la matriz de confusión de la Tabla 1, se aprecia que, si bien estamos alejados de clasificar la totalidad de las clases ampliamente minoritarias, se puede afirmar que logramos ascender a un 50% en una clasificación correcta de varias de ellas. Esto facilita, en gran medida, la realización de una predicción adecuada de la tendencia detectada en este área.

Una vez concluido el aprendizaje, se pueden analizar nuevos casos. Para ello, se introducen datos de los atributos considerados en la muestra, la que sería ampliamente mejorada si se solicitaran solamente aquellos atributos que el árbol de decisión consideró relevantes o, mejor aún, se pudiera realizar una especie de asistente (*wizard*) que transite sobre el camino determinado por el árbol de decisión. Así, se podrán ajustar al mínimo los datos solicitados para una predicción, sin importar la amplitud que pudiera tener la muestra en forma horizontal.

Realizada la carga de un nuevo caso a predecir, se realizó el procesamiento basado en el aprendizaje conseguido con el algoritmo y, una vez obtenido el resultado de la predicción que se realiza en segundos, se concluye con la clasificación obtenida.

5 Conclusiones

Se ha logrado desarrollar una aplicación en Java, que puede prestar servicios a una HCE, tanto para servir de apoyo en la predicción de diagnósticos, como así también en casos de prevención.

Si bien los datos que se procesaron requirieron un amplio debate de manera interdisciplinaria con profesionales de la salud que colaboraron con el trabajo, fue necesario un preprocesamiento intenso. Luego de varios intentos de extraer información de utilidad, podemos afirmar que fueron ampliamente satisfactorios los

resultados alcanzados, ya que se demostró la posibilidad de incorporar este desarrollo como *Sistema de Soporte a las Decisiones Clínicas* (CDSS por sus siglas en inglés).

Es necesario resaltar que, evidentemente, la calidad de la predicción obtenida recae sobre la información que se puede tomar como insumo; asimismo, es oportuno enfatizar que la información utilizada en el caso de estudio, necesariamente requerirá adicionar el cruzamiento con resultados de análisis clínicos, imagenología, anatomía patológica, así como también otros datos provenientes de diferentes eventos médicos. De este modo, se podrá aportar mejor calidad a la muestra y manejar datos discretizados y no tan polarizados como los de la muestra considerada.

Es destacable el valor que está adquiriendo en el sector sanitario el tratamiento de los grandes volúmenes de datos y el rol de los mismos como fuente estratégica, la que permitirá avanzar en la toma de decisiones clínicas. Finalmente, concluimos que, a través de las tecnologías de minería de datos, los profesionales de la salud tienen a su alcance herramientas que pueden ser utilizadas como apoyo en la toma de decisiones médicas vinculadas al diagnóstico clínico.

Agradecimientos.

Este trabajo está parcialmente financiado por los proyectos de Investigación *"Estudio Comparativo y Análisis de Rendimiento de los Lenguajes de Manipulación de Datos en Bases de Datos Orientadas a Objetos y Bases de Datos Objeto-Relacionales"*, PID 7042 (UNER), y *"Guías para aplicación de Normas de Calidad para los procesos de Ingeniería de Software en productos desarrollados con Lenguajes de Programación Open Source: relevamiento y aplicación en PYMES de la zona de influencia de la UNER Concordia"*, PID 7049 (UNER).

Referencias.

1. Balas, E. A., Vernon, M., Magrabi, F., Gordon, L. T., Sexton, J.: Big Data Clinical Research: Validity, Ethics, and Regulation. In MEDINFO 2015: EHealth-enabled Health: Proceedings of the 15th World Congress on Health and Biomedical Informatics, Vol. 216. IOS Press (2015) 448
2. Molina, J., García, J.: Técnicas de análisis de datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka. Universidad Carlos III de Madrid España (2006)
3. Hand, D. J., Mannila, H., Smyth, P.: Principles of Data Mining. MIT press (2001) ISBN: 026208290x.
4. Witten, I. H., Frank, E., Hall, M. A.: Data mining: Practical machine learning tools and techniques (3a. ed. --.). s.l.: Elsevier (2011)
5. Wasan, S. K., Bhatnagar, V., Kaur, H.: The impact of data mining techniques on medical diagnostics. Data Science Journal, Vol. 5. (2006) 119-126.
6. Milovic, B., Milovic, M.: Prediction and Decision Making in Health Care using Data Mining. International Journal of Public Health Science (IJPHS), Vol. 1, Nº 2. (2012) 69-78 ISSN: 2252-8806.

7. Zamarrón Sanz, C., García Paz, V., Calvo Álvarez, U., Pichel Guerrero, F., Rodríguez Suárez, J.: Aplicación de la Minería de Datos al estudio de las alteraciones respiratorias durante el sueño. Revista Pneuma, Vol. 6. (2006) 156-166
8. Nayak, A., Poriya, A., Poojary, D.: Type of NOSQL databases and its comparison with relational databases. International Journal of Applied Information Systems, Vol. 5, Nº 4. (2013) 16-19
9. Mehta, R. G., Mistry, N. J., Raghuvanshi, M.: Impact of Column-oriented Databases on Data Mining Algorithms. International Journal of Advanced Research in Computer and Communication Engineering (2013)
10. Carabio, A. L. R., Benedetto, M. G., Falappa, M. A.: Comportamiento de Bases de Datos No Relacionales en Entornos Distribuidos. In XVIII Workshop de Investigadores en Ciencias de la Computación (2016)
11. Java Platform, Standard Edition (Java SE) 8, http://docs.oracle.com/javase/8/index.html
12. Bouckaert, R.R., Frank, E., Hall, M.A., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: WEKA Manual for Version 3-6-12. University of Waikato (2014)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Vol.11, Issue 1. (2009)
14. Kaur, G., Chhabra, A.: Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, Vol. 98, Nº 22. (2014)
15. Martínez, E. H., Sanjurjo, R. L.: Minera de datos aplicada a la detección de Cáncer de Mama (2009)
16. Shafique, U., Majeed, F., Qaiser, H., Mustafa, I. U.: Data Mining in Healthcare for Heart Diseases. International Journal of Innovation and Applied Studies, Vol. 10, Nº 4. (2015) 1312
17. Mitchell, T. M.: Machine Learning. McGraw-Hill Science/Engineering/Math (1997)